Stimulus-stimulus Association via Reinforcement Learning in Spiking Neural Network

Nooraini Yusoff and Farzana Kabir Ahmad
School of Computing, College of Arts and Sciences
Universiti Utara Malaysia, 06010 UUM Sintok
Kedah, Malaysia
{nooraini, farzana58}@uum.edu.my

***Abstract*—In this paper, we propose an algorithm that performs stimulus-stimulus association via reinforcement learning. In particular, we develop a recurrent network with dynamic properties of Izhikevich spiking neuron model and train the network to associate a stimulus pair using reward modulated spike-time dependent plasticity. The learning algorithm associates a prime stimulus, known as the predictor, with a second stimulus, known as the choice, comes after an inter-stimulus interval. The influence of the prime stimulus on the neural response after the onset of the later stimulus is then observed. A series of probe trials resemble the retrospective and prospective activities in human response processing.**

Keywords-component; associative learning, spiking neural network, reinforcement learning, spike-time dependent plasticity, priming effect

I. INTRODUCTION

In some environments, the effect of an earlier seen stimulus could influence the processing of response for a later stimulus. This is what has been termed as the effect of priming in memory recall. The cognitive behaviour of priming effect shows signs of influence of previous information on the perception of subsequent information [10]. The effect is a result of 'spread activation' mechanism in the brain in which a recently probed stimulus invokes its associated information, consequently strengthening the retrieval of information of a later proceeded stimulus when both are related. For this case, the prime stimulus acts as a cue for the later stimulus.

To support this, recordings in the associative cortex of a group of monkeys while performing a set of visual discrimination tasks as published by [9] have shown persistently increasing activity in the brain when presented with a stimulus. The activity was indicating not only response to the shown stimuli but also the stimuli that the monkeys were expecting to be seen, one that had been associated to the shown stimuli.

From past studies of spiking neural networks (SNNs), we have found that little work has been reported on its implementation in reinforcement learning paradigm. What we found typically are abstract algorithms not based on explicit neural modeling. Only recently there seems to be increasing work on modelling of reinforcement learning in SNN, after neurophysiological data linking dopamine signals in the brain that is believed to play an important role in enhancement of synaptic changes [7], [13]. The dopamine signals are hypothesised to be responsible for the reward acquisition mechanism in the brain, thus giving us some clue on connection between synaptic plasticity at the microscopic level with behavioural changes in animals.

In reinforcement learning [17], agents must update their internal parameters in order to maximise reward over time at a given task. This is implemented through a series of trial-and-error action-rewards in response to environmental stimuli. Unlike supervised and unsupervised approaches, where in most cases learning follows some specific rules with given initial state, in the reinforcement approach, agents explore and exploit their unknown identity states to establish a learning policy.

In this paper, inspired by the realistic properties of the artificial network with spiking neuron and the interesting behaviour of priming effect, we propose a stimulus-stimulus association in a reinforcement learning paradigm. Here, we hypothesise that, in an environment with rich and realistic dynamics, a network could be simulated to exhibit retrospective and prospective activity in the brain (priming effect). In particular, we extend our work in [18] by simulating the priming effect in Colour-Word stroop test [15] setting.

II. LEARNING IMPLEMENTATION

A. Spiking Neural Network Model

The neural network simulation model consists of 800 excitatory and 200 inhibitory spiking neurons. The spiking properties of each neuron are modelled with Izhikevich spiking neuron model [3], [4]. To simulate the properties of real cortical neurons, all excitatory neurons exhibit regular spiking type neurons and all inhibitory neurons are fast spiking neurons.

The connectivity between neurons is random and sparse with probability $p=0.1$. Each excitatory neuron is randomly connected to 100 neurons, and each inhibitory neuron is randomly connected to 100 excitatory neurons. The connectivity between neurons is modelled after the structure

of neocortex with random and sparse connectivity [6]. Each synaptic connection, from neuron i to j, is defined by two parameters: a weight w_{ij} (in the range of $0 < w \leq 4.0$ mV) and a synaptic transmission delay d_{ij} (between 1 to 20 ms, randomly distributed, [4]. Weights are initialised with 1.0 and -1.0 for excitatory and inhibitory weights, respectively. Synaptic plasticity is only applied on excitatory weights whilst the inhibitory weights are not plastic.

The excitatory population is divided into subpopulations of neurons namely m stimulus groups S, n response groups R and non-selective neurons NS. In our model the inhibitory subpopulation IH acts as global inhibition (Fig. 1). Each stimulus group consists of 50 neurons and each response group consists of 100 neurons.

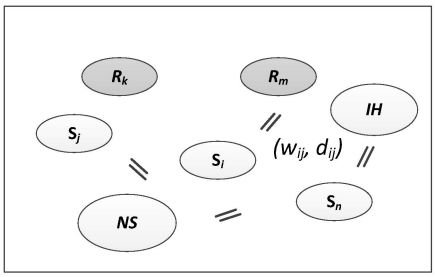

Figure 1. Schematic view of a recurrent spiking neural network consisting of 80% excitatory (N_E) neurons and 20% of inhibitory (N_I) neurons, with sparse and random connectivity, p = 0.1 (no self-feedback), i.e. $N_E \rightarrow \{N_E, N_I\}$ and $N_I \rightarrow N_E$ (see above text for details). Each synaptic transmission has random delay $d \in [1, 20]$ ms. Neurons are divided into subpopulations of stimulus groups (S), response groups (R), non-selective neurons (NS) and inhibitory pool (IH). S and R are composed of 50, and 100 excitatory neurons, respectively.

B. Associative Learning via Reinforcement Approach

In every learning simulation, we initiate a network with random activity, for the first 100 ms. For this purpose, we stimulate an arbitrary neuron with 20-pA current for every ms. With the same random activity as the background, the network is given a set of pair-response mappings $(S_i,S_j) \rightarrow R_k$, with different pairing strategies depending on the task. For each learning trial, at time t_n we present the first stimulus, i.e S_i to the network by stimulating all neurons in S_i with a strong current of 20 pA. After an inter-stimulus interval (ISI), we stimulate all neurons with the same amount of current to the second stimulus, i.e. S_j to be associated to S_i. An optimal ISI is chosen from a range of 10 – 50 ms based on a preliminary experiment. Each learning runs for 20 minutes simulated time with random presentation of stimulus pairs.

Within a 20-ms time window from the onset of the second stimulus, we count the number of activations in the response groups, i.e, R_k. The response group with the highest number of activations is considered to be the winner. To accelerate the learning, some bias current is supplied to the target winner. This is implemented via stimulation of 20-pA current to arbitrary neurons (with probability of neurons to be selected is between p=0.25 to 0.5, weak to strong potentiation) in the target response group. The next learning pair is presented after a 100-ms delay from the offset of each response interval.

Synapse reinforcement is implemented based on a reward policy. The network is positively rewarded if a target response group is the winner of a learning trial, or otherwise negatively rewarded. The reward policy determines the amount of synapse potentiation (i.e. strong or weak potentiation) or depression.

In our model, learning is implemented onto excitatory synapses only for every 10-ms time step. The synaptic efficacy is dependent on a reinforcement signal (i.e. reward signal), *r(t)*, derived from a reward policy. The signal modulates the synaptic changes read from a spike-timing dependent plasticity (STDP) function (as in 4).

$$\Delta w_{STDP} = \Theta \{A_+ e^{-\Delta t/\tau_+}, \Delta t \geq 0; A_- e^{\Delta t/\tau_-}, \Delta t < 0\} \quad (1)$$

From (1), the synapse is potentiated if the difference in firing times (Δt) between a postsynaptic neuron and its presynaptic neuron (i.e. t_{post}-t_{pre}) is ≥ 0, otherwise the synapse is depreciated. The magnitude of potentiatiaon (depression) is given by $A_+ e^{-\Delta t/\tau_+}$ ($A_- e^{\Delta t/\tau_-}$), where A represents the maximal change when the spike timing difference Δt is approaching 0, and τ is the time constant (in ms). For our STDP curve, $\tau_+ = \tau_- = 20$ ms, $A_+ = 0.1$, and $A_- = 0.15$ (following [2]).

The reinforcement signal *r(t)* is obtained from a reward policy (in Table 1) that is based on the number of neuron firings (F) of response groups within a response interval of 20 ms. The signal determines the amount of modulation to the summation of Δw_{stdp}. Therefore, the reward modulated STDP learning holds [2], [16]:

$$\Delta w(t) = [\alpha + r(t)]\, z(t) \quad (2)$$

where α is the activity-independent increase of synaptic weight, *r(t)* and *z(t)* are the reinforcement signal (5) and the eligibility trace, respectively. *z(t)* represents the summation of Δw_{stdp} obtained from (1).

TABLE I. REWARD POLICY

Type of reward	**Reinforcement signal, *r(t)***	**Response firing rate, *F***
strong +ve reward	$r(t-1) + 0.5$	$F_i \geq 2F_j$
weak +ve reward	$1 - F_j/F_i$	$F_i > F_j$
-ve reward	-0.1	$F_i < F_j$

a. F_i and F_j are the number of firings of a target response group and non-target response group, respectively.

III. SIMULATION RESULTS

We trained a network first to learn a simple learning task whose aim was to select the optimal parameters for further use in other extended pair-associate tasks. For this purpose, the network was trained to associate 4 paired predictor-choices, each of which was reinforced to respond to one of 2 target responses, R_A or R_B. We studied the network dynamics by observing the interaction between 2 stimulus groups of the paired predictor and choice with R_A or R_B. The learning pairs were as follows: $\{(S_0,S_1) \rightarrow R_A, (S_2,S_3) \rightarrow R_B, (S_4,S_5) \rightarrow R_A, (S_6,S_7) \rightarrow R_B\}$.

During the training, at stimulation times t_n and t_{n+ISI}, all neurons (N=50) in the predictor and choice groups were supplied with a superthreshold current of 20 pA. An example of spike raster plot at the early phase of learning and after a number of trials is depicted in Fig. 2.

At the early phase of learning, activation in each stimulus group was only evoked by coincident firings of 50 stimulated neurons. After a number of rewards given based on the activation rate of the response groups (within an interval of 20 ms from the onset of a choice), the network response to the paired stimuli became reinforced. The synaptic connections from the paired predictor-choice were stronger compared from other non-reinforced stimulus groups. Hence, neurons in paired stimuli could strongly influence their postsynaptic targets.

A. *Associative Memory Recalls*

- Noise tolerance

To test for noise robustness in a response recall, for every learned stimulus pair, we performed response recalls by randomly activating neurons in the predictor and choice groups. We tested a trained network with selectivity of neurons to be stimulated, p_n, from 0.5 to 1.0.

The averaged recall performance over 100 probe trials showed that the each stimulus group required minimal activations of 70% (35 out of 50) of neurons at minimum of 65.48% of correct recalls. In other words, a network with random synaptic connectivity of 0.1, tolerates maximal distortion probability of 0.3.

- Recalls with only choice groups (control condition), congruent pairs, and incongruent pairs.

We ran a series of probe trials to see the effect of priming in response recalls. The trained network was probed with 3 conditions of stimuli namely control – the network was only presented with learned choices without their predictor, $\{S_1, S_3, S_5, S_7\}$, congruent – learned paired stimuli (predictor-choice), $\{(S_0,S_1), (S_2,S_3), (S_4,S_5), (S_6,S_7)\}$, and incongruent – predictor and choice with conflicting responses, $\{(S_0,S_3), (S_2,S_1), (S_4,S_7), (S_6,S_5)\}$. For trials with congruent and incongruent conditions, the ISI was 15 ms.

In response to a single stimulus (control condition – with choice only), the averaged performance over 100 trials was 53.93%. When presented with congruent pairs, the percentage of correct recalls achieved 95.85%, meanwhile averaged correct recalls when responding to incongruent pairs decreased to 42.28%. This indicates a facilitation effect when a choice is preceded with its correct paired predictor. Priming the network with a predictor acting as a cue to its choice gives advantage in recalling the response. Meanwhile high competition or interference exists when the network is probed with predictor-choice having conflicting target responses.

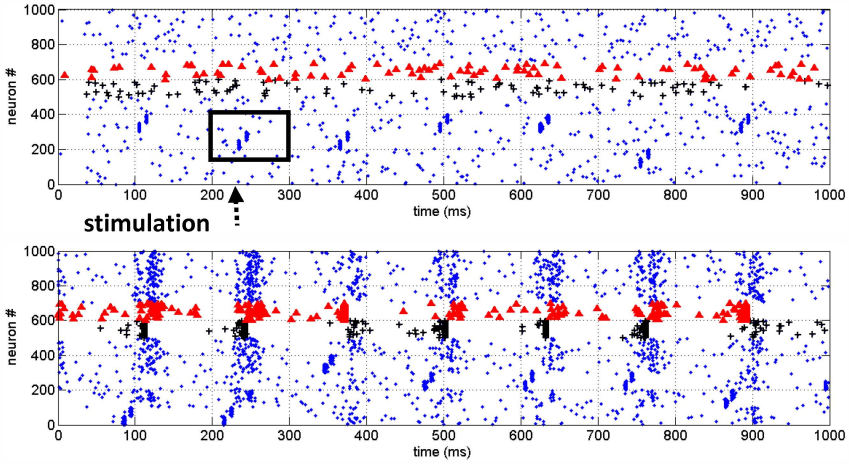

Figure 2. Spike raster plot of a learning (top) at the early phase, and (bottom) after 500 seconds, within 1000 ms time window. Neurons in the response groups are marked with '+' (neurons 501-600) and '^' (neurons 601-700).

B. *Simulation of Cognitive Priming Effect in Associative Memory Recall*

- Colour-Word stroop paradigm

For initial understanding on the dynamics of information processing in the priming effect, a cognitive behaviour known as the Colour-Word Stroop effect [15] has been chosen for our experimental paradigm. In particular, we studied cognitive control in an environment that can cause automatic responses to certain classes of stimuli. These stimuli are the almost unavoidable negative cues that would result in distraction from the intended task due to one area of the brain dominating and suppressing the response of other functional areas. In such a condition, higher inhibition is required to boost attentional resources to provide the target response to the goal, e.g., [5], [11]. The goal of this initial stage of our work was to study memory formation and retrieval in terms of interference and facilitation.

In a real Stroop test, subjects are required to respond (verbally) to a sequence of stimuli. The stimuli are the coloured colour-words in three conditions; control (e.g. a non coloured colour-word – e.g. RED written in black or a non colour-word – e.g. BOOK written in red), congruent (e.g. a colour-word RED written in red) and conflicting (e.g. a colour-word RED written in green). Participants are asked either to read the colour-words or to name their colours while the reaction times of performing the task are observed.

The findings from Stroop studies concluded that there is increased reaction time in naming the colour of the printed colour-word denoting a different colour, while the subjects could easily read the word and ignore the colour. Meanwhile the congruence of the word and its colour reduces the time of response processing to the colour name. The automatic processing of a stimulus (here: word meaning) can cause interference or facilitation to other stimuli when this stimulus is presented in temporal or spatial proximity to the target stimulus (here: print colour) and can be seen as a type of priming effect. The priming effect occurs due to the automatic (and fast) processing of irrelevant stimuli influencing less automated (and slow) target stimulus processing. The prime stimulus can be a cue if it is congruent with the target stimulus and facilitate its response; otherwise it is a distractor if they are conflicting. It is believed that whenever the interference occurs, cognitive inhibition is performed actively, requiring higher attentional control in producing the desired response to the intended stimulus.

- Stroop stimuli

We develop a recurrent network with 1000 Izhikevich spiking neurons composed of 800 excitatory and 200 inhibitory neurons. The connectivity between neurons is sparse and random with probability of 0.1 as described in Section II (A). For Stroop simulation, from the excitatory neurons pool, there are 2 stimulus groups to represent word stimuli, e.g. "RED" and "GREEN", 2 stimulus groups that are selective to colour stimuli, e,g. "red" and "green", and there are 2 response groups to represent respectively the "Red" and "Green" colour concept (Fig. 3).

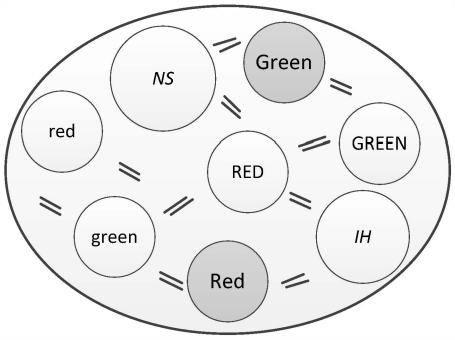

Figure 3. A sparse and random recurrent network for Stroop paradigm. "RED" and "GREEN" are selective to word stimuli whilst "red" and "green" represent colour stimuli, "Red" and "Green" are the response groups, *NS* and *IH* are respectively the non-selective and inhibitory pool.

- Stimulus presentation

We first train a network to learn individual stimulus groups, i.e. word stimuli - "RED", "GREEN" and colour stimuli – "red" and "green". The stimulation is run in the following way; each stimulus group is partitioned into two parts, namely a predictor and a choice separated with 15-ms delay. For example a neuronal word group "RED" is first stimulated via its predictor followed by the choice (Fig. 4A). At learning stage, stimulations to all predictor stimuli are implemented randomly with probability of 0.5 for each neuron to be induced with current. This is to simulate an initial percept of word or colour stimulus (e.g. early stage of attention). Stimulations to choice stimuli in which the responses are counted are supplied with currents to all of their neurons. The response is based on the spike count in groups representing the colour concept, i.e. "Red" and "Green" within 20 ms interval from the onset of a choice stimulus. The group with the greatest number of spike counts is chosen to be the winner.

In learning with Stroop stimuli, the word stimuli are presented randomly three times more than the colour stimuli. This is based on a theoretical account of the Stroop effect that conjectures a response for a word is automatic as a result of effect of practice that in human, words are over-learned, consequently to create habitual response [12], [14]. The well-known classic Stroop effect model by [8] has also incorporated this account into their learning. In their parallel distributed processing model trained with backprogation learning, word stimuli were presented 10 times more than colour stimuli. Hence, this could result stronger processing path (connection weights) for word stimuli. For our simulation experiment, we ran the training experiment for 20 mins simulation time repeated for 10 different network configurations.

- Probe trials

In Stroop probe trials, there are three types of stimuli namely control, congruent and incongruent. The Stroop tasks, for word reading or colour naming, are distinguished according to the sequence of predictor and choice. The target task is determined by the choice stimulus group. For example if the task is word reading, the choice stimulus is either the word "RED" or "GREEN", meanwhile the choice stimulus is the colour "red" or "green" if the task is colour naming. Here we study the effect of a predictor stimulus on its choice stimulus.

For control stimuli, a trained network is tested with learned word and colour stimuli. Each group is first triggered via their predictor that we randomly supply superthershold current of 20 pA to selected neurons with minimum p = 0.5. To minimise the variability in firing activity and only study the Stroop effect, all neurons in every choice group are activated with external currents. The stimulation to a predictor and its choice is delayed for 15 ms. A similar procedure of stimulation is applied for congruent (e.g., "RED" and "red") and incongruent (e.g., "RED" and "green") Stroop stimuli. For Stroop tests with congruent and incongruent pairs, the predictor and choice of the same group (e.g. "RED") are stimulated concurrently using the same method of stimulation for control groups but without delay. Meanwhile stimulation to different groups is implemented with delay (i.e. 15 ms), see Fig. 4. The testing result shows the averaged percentage of performance over a number of trials, i.e. performance = (number of correct recall/number of trials)*100.

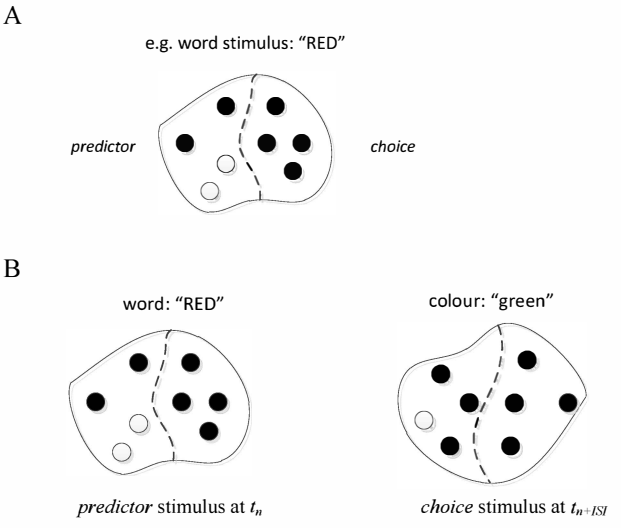

Figure 4. (A) Stimulation to a stimulus group is implemented via delivery of current to its predictor, randomly selected, followed by stimulation to all neurons to its choice group with delay of 15 ms during training. (B) Stimulation procedure for incongruent (also for congruent) Stroop stimuli. The same stimulation method as in A is applied for triggering neurons in the same group, e.g. word stimulus "RED", but with simultaneous activation of its predictor and choice. Stimulation to different groups is delayed with inter-stimulus interval ISI. For congruent pairs, ISI is fixed with 15-ms delay.

- Stroop simulation results

Fig. 5 shows the Stroop test performance for ISI = 15 ms for word reading (WR) and colour naming (CN) tasks. The effect of over-learned word stimuli can be seen in performance of control stimuli with 72.2% and 63.4% for WR and CN respectively. The facilitation of word stimuli to colour stimuli was achieved with 99.9% and 99.40% of correct recalls for both WR and CN, respectively, with congruent pairs. Meanwhile, the interference effect of undesired word responses to colour stimuli in CN was shown in recalls with incongruent pair. The dominance of word stimuli can be observed from less negative effect for WR and greater effect in CN. Even with 15 ms delay, the correct response recall rate for CN was only 30.9% (WR: 65.8%).

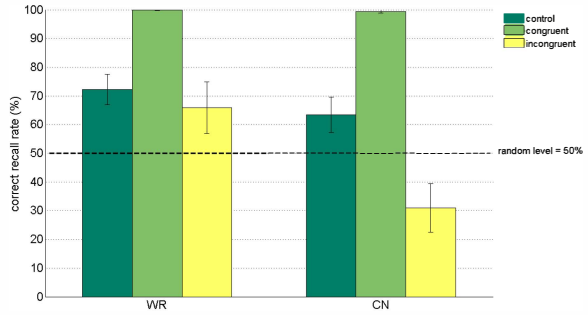

Figure 5. Stroop performance for control, congruent and incongruent in probe trials with ISI = 15 ms.

The results are consistent with the Stroop behavioural study in human by [1] (as in Fig. 6). Even though in latter, the key measurement is the response time, the delay in response processing is a result of cognitive inhibition due to conflicting responses and error detection that affect accuracy (a detailed review on the role of anterior cingulate cortex in conflict and error detection for response reaction time and accuracy can be found in [5]).

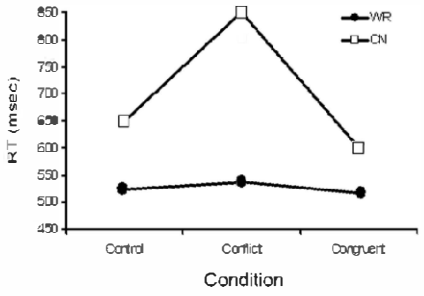

Figure 6. Performance results for Stroop task from the empirical study, reproduced from [1].

IV. CONCLUSION

In this study, we implement our proposed stimulus-stimulus association learning to simulate the cognitive priming effect. In particular, we simulate the colour-word stroop effect exhibiting the interference and facilitation in response processing. Interestingly, in our experiments, despite of the chaotic spiking network resulting from random activities and synaptic delays, the learning rewarding mechanism could still reinforce the right synapses even on delayed paired stimuli. It has been shown that, for a paired predictor-choice, priming a network with the predictor as a cue facilitates recalls to correct response.

In cognitive modelling of the Stroop effect, we show that the cognitive behaviour can be modelled with a generic neural network architecture. In some existing Stroop models (e.g. [5], [8]), a specific network architecture is required to simulate the automaticity of word stimuli. In ours we only include a learning mechanism that could explain a negative priming effect in information processing. Our model is relevant and can explain the behaviour according to some theoretical accounts [12]. This demonstrates that the model can be used to study human cognition related to priming effect. Therefore this could offer a support tool for understanding and diagnosis of certain cognitive impairment. Furthermore, we have also successfully implemented the proposed learning scheme in a task of visual recognition of real images [19]. We can see some potential applications that can be built inspired by the cognitive process of priming effect. For example, an agent can be trained to associate stimuli for visual recognition task, path tracking, and multimodal authentication e.g. audio-visual.

Despite the model biological realism and simple computation, the model could be improved to capture reaction time in response processing. This we believe is a trade-off between synaptic transmission delays in our spiking network. With the absence of delays, we can measure response time when some firing rate exceeds a threshold at certain time t. Nevertheless, this could only be possible if a neuron only acts as an integrator with fixed neuron firing threshold. In our network with Izhikevich spiking neuron, as with biological neurons, a neuron does not have a fixed threshold.

ACKNOWLEDGMENT

This research has been funded by the Ministry of Higher Education (Malaysia).

REFERENCES

[1] K. Dunbar and C.M MacLeod, "A Horse Race of a Different Color: Stroop Interference Patterns with Transformed Words," Journal of Experimental Psychology: Human Perception and Performance, vol. 10, 1984, pp. 622-639.

[2] E. M. Izhikevich, "Solving the distal reward problem through linkage of STDP and dopamine signaling," Cereb Cortex, vol. 17, 2007, pp. 2443–2452.

[3] E.M Izhikevich, "Simple Model of Spiking Neurons," IEEE Trans. Neural Networks, 14(6), pp. 1569-1572, 2003.

[4] E. M. Izhikevich, "Polychronization: Computation with Spikes," Neural Computation, vol. 18, 2006, pp. 245-282.

[5] G.B. Kaplan, N. S. Sengor, S. Gurvit and C. Guzelis, "Modelling The Stroop Effect: A Connectionist Approach," Neurocomputing, vol. 70, 2007, pp. 1414-1423.

[6] V. Cutsuridis and T. Wennekers, "Hippocampus, microcircuits and associative memory," Neural Networks, vol. 22, 2009, pp. 1120-1128.

[7] Q. Gu, "Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity," Neuroscience, vol. 111, 2002, pp. 815-835.

[8] J. D. Cohen, K. Dunbar and J. L. McClelland, "On the Control of Automatic Processes: A Parallel Distributed Account of the Stroop Effect," Psychological Review, 97(3), 1990, pp. 332-361.

[9] C. A. Erickson and R. Desimone, "Responses of Macaque Perirhinal Neurons during and after Visual Stimulus Association Learning," Journal of Neuroscience, 19(23), 1999, pp. 10404-10416.

[10] M. G. Filippova, "Does Unconscious Information Affect Cognitive Activity?: A Study Using Experimental Priming," The Spanish Journal of Psychology, 14(1), 2011, pp. 20-36.

[11] S. A. Herd, M. T. Banich and R. C. O'Really, "Neural Mechanism of Cognitive Control: An Integrative Model of Stroop Task Perfromance and fMRI Data," Journal of Cognitive Neuroscience, 18(1), 2006, pp. 22-32.

[12] C. M. MacLeod and K. Dunbar, "Training and Stroop-like interference: Evidence for a continuum of automaticity," Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 14, 1988, pp. 126-135.

[13] W. B Smith, S. R. Starck, R. W. Roberts and E. M. Schuman, "Dopaminergic Stimulation of Local Protein Synthesis Enhances Surface Expression of GluR1 and Synaptic Transmission in Hippocampal Neurons," Neuron, vol. 45, 2005, pp. 765-779.

[14] R. M. Shiffrin and W. Schneider, "Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory," Psychological Review, vol. 84, 1977, pp. 127-190.

[15] J. R. Stroop, "Studies of Interference in Serial Verbal Reactions," Journal of Experimental Psychology, vol. 18, 1935, pp. 643-662.

[16] R. Legenstein, D. Pecevski and W. A. Maass, "Learning Theory for Reward-Modulated Spike-Timing-Dependent Plasticity with Application to Biofeedback," PLoS Computational Biology, 4(10), 2008, pp. 1-27.

[17] R. S. Sutton and A. G. Barto, Reinforcement learning: an introduction. Cambridge (MA): The MIT Press, 1998.

[18] N. Yusoff and A. Gruning, "Learning Anticipation through Priming in Spatio-temporal Neural Networks," Proc. Neural Information Processing, LNCS 7663, 2012.

[19] N. Yusoff and A. Gruning, "Biologically Inspired Temporal Sequence Learning," Journal of Procedia Engineering, vol. 41, 2012, pp. 319-325.